SCHOLARS
STRATEGY NETWORK

How Test Scoring Decisions Affect School Effectiveness Ratings

Joshua Baer Gilbert, Harvard University

Which schools most improve student learning? This question is critical for parents, educators, and policymakers. For state education agencies looking to make decisions about school accountability and resource allocation, having reliable information on school effectiveness is essential. One approach to answering this question would be to identify the schools with the highest average end-of-year test scores. However, this approach is likely to be misleading, because students *enter* schools with different levels of preparation. To address this issue, some states rank school effectiveness using a statistical approach called a *Value-added Model* or VAM.

VAMs statistically adjust end-of-year test scores for differences in prior test scores and student characteristics like race, gender, and socio-economic status. Schools that have high VAM rankings do *better than expected* given prior student scores and demographics. To make the logic of VAMs more concrete, imagine two middle schools with equal 6th grade average scores. If the students in School A have high 5th grade math scores, and students in School B have low 5th grade math scores, School B would have a higher VAM ranking than School A because students in School B entered 6th grade with lower scores. As of 2018, 15 states use VAMs in their accountability systems, so the information VAMs provide has significant implications for how schools in these states are compared and evaluated.

When Test Scoring Methods Change, So Do School Rankings

A new study sheds light on a challenge with VAMs: how the tests are scored. Historically, many tests were scored by counting up the number of correct answers on a test—one student might score 30/40, another 35/40. However, most modern testing systems use a statistical approach to score tests, in which test items are weighted differently and better account for how tests change over time. The statistical scoring approach raises an important question: how would a school's ranking change had a different statistical model been used to generate the test scores? This is not a purely theoretical issue. For example, Massachusetts and New York use different statistical models to produce student test scores, yet it remains unclear how much such statistical test scoring decisions matter in practice.

The study tested this issue by using 18 datasets to see how much school VAM rankings shifted when the statistical scoring model changed. Importantly, the students, schools, and test items were held constant here: only the statistical model used to score the test varied. The study found that in many datasets, over half of the schools were ranked in different quartiles (e.g., the bottom 25% of schools). In one case, over 30% of schools were ranked in the "bottom 10%" when only the scoring method changed. Low VAM rankings can have important real-world consequences. For example, the DC IMPACT program used VAMs to identify low performing teachers and terminate them if they did not improve. Students can also be affected by these decisions: one study found that low VAM rankings predicted higher teacher turnover.

How State Education Agencies Can Make School Accountability Systems More Reliable

What should state education agencies do about this hidden source of variation in school accountability rankings? We have two recommendations.

1. First, make the statistical scoring model transparent. While the scoring models are generally publicly available, they tend to be buried in long technical reports. Making the scoring approach more prominent in test documentation or state websites and justifying the scoring approach in plain language will likely aid public understanding of test scores and increase trust in school accountability policy.
2. Second, whenever VAM rankings are used for accountability, calculate the rankings using multiple scoring methods. Calculating rankings across multiple methods (and potentially averaging them using modern statistical methods) ensures that a very high or low performing school is not just getting "lucky" with one scoring system.

In sum, VAMs can be useful tools to identify which schools are helping students learn. However, technical aspects of VAMs such as which statistical model to use to score tests have underappreciated consequences on which schools are identified as "effective." By making scoring decisions more transparent, state education agencies can help policymakers make better use of test score data for school accountability.

Read more in Gilbert, J.B., Soland, J.G. and Domingue, B.W. (2026), The Sensitivity of Value-Added Estimates to Test Scoring Decisions. Educational Measurement: Issues and Practice, 45: e70011. https://doi.org/10.1111/emip.70011